BUNDESREPUBLIK DEUTSCHLAND

Bescheinigung

Die Siemens Aktiengesellschaft in München/Deutschland hat eine Patentanmeldung unter der Bezeichnung

"Such- und Navigationseinrichtung für Hypertext-Dokumente"

am 9. März 1999 beim Deutschen Patent- und Markenamt eingereicht.

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

Die Anmeldung hat im Deutschen Patent- und Markenamt vorläufig das Symbol G 06 F 17/30 der Internationalen Patentklassifikation erhalten.

München, den 15. April 1999

Deutsches Patent- und Markenamt

Der Präsident

Im Auftrag

Ebert

Aktenzeichen: 199 10 357.7

Such- und Navigationseinrichtung für Hypertext-Dokumente

Technisches Gebiet

Die Erfindung betrifft die Navigation und Suche in durch Verweise verketteten Dokumente, die überwiegend als Hypertext-Dokumente bezeichnet werden.

Stand der Technik

Mit Verweisen verkettete Dokumente werden üblicherweise als Hypertext-Dokumente bezeichnet. Dabei sind die im Internet verwendeten, in der "Hypertext Markup Langauage" (HTML) beschriebenen Dokumente weit verbreitet und allgemein bekannt. Ein weiteres Beispiel für Hypertext-Dokumente sind die in den von der Firma Microsoft vertriebene graphischen Benutzeroberfläche "WINDOWS" enthaltenen Hilfedateien. Im folgenden sollen HTML-Seiten als stellvertretend für alle Hypertext-Dokumente angesehen werden.

Wenngleich die Navigation über die Hyperlinks eines Hypertext-Dokuments die Suche nach Information wesentlich gegenüber einem traditionellen Dokument mit hierarchischer Kapitelstruktur wesentlich verbessert hat, so sind doch weitere Hilfsmittel zur Suche und Navigation notwendig. Dazu gehört sicherlich ein Index, der von Suchwörtern zu den entsprechenden Seiten verweist.

Als weiteres Hilfsmittel sind "Suchmaschinen" bekannt. Diese werden mit einem oder mehreren Stichwörtern aufgerufen, die auf einen im vorarb erstellten, kontinuierlich aktualisierten, meist sehr umfangereichen, nicht direkt sichtbaren Index angewendet werden und Verweise auf eine Anzahl von

Dokumenten, in denen diese Stichworte erwähnt sind, anzeigen. Dabei werden bei der Erstellung dieser Indizes bei HTML-Dokumenten entweder nur die über das META-Tag angebbaren Schlüsselwörter verwendet, oder es werden zusätzlich die Text-Inhalte weiterer Tags, insbesonder des "TITLE"-Tags, oder zusätzlich der gesamte Textinhalt verwendet. Dies ist primär eine Frage des zu indizierenden Datenbestandes in Relation zu den verfügbaren Betriebsmitteln.

Allerdings ist bei den Suchmaschinen zum einen die richtige Wahl der Suchwörter ausschlaggebend für ein gutes Suchergebnis. Zum zweiten wird der Zusammenhang der Dokumente untereinander weder berücksichtigt noch dargestellt.

Es tritt aber in der Praxis häufig der Fall ein, daß man bereits eine halbwegs passende Hypertext-Seite gefunden hat, die jedoch noch nicht die gewüschte Information enthält. Man muß also die Verweise systematisch vor- und zurück absuchen und die Hypertext-Seiten selbst inspizieren, um die gewünschte Information zu finden.

Hypertext-Seiten stellen in ihrer Basis-Struktur einen Baum dar, weil jede Seite als Knoten mit Verweisen zu untergeordneten Knoten erscheint. Die Rück- und Querverweise jedoch stören diese Struktur empfindlich. Dennoch ist es als Navigationshilfe bekannt, einen Strukturbaum der Hypertext-Dokumente anzuzeigen, der auch als 'site-map' bezeichnet wird. Hierbei wird, ausgehend von einer, meist als 'home page' bezeichneten, Wurzel ein Baum aufgebaut, wobei alle der Baumstruktur widersprecheden Verweise ganz unterdrückt oder nur schwach angezeigt werden. Hierzu sind eine Anzahl von meist zweidimensionalen graphischen Darstellungsformen bekannt. Neuerdings werden auch dreidimensionale Abbildungen gewählt, die der Benutzer interaktiv im Raume drehen

kann, wobei eine entsprechende Projektion auf eine zweidimensionale Fläche angezeigt wird.

Nachteilig ist dabei, daß diese Darstellung nur mit einem kurzen Text, meist dem vergebene Titel, bezeichnet sind. Damit ist zwar die Navigation übersichtlicher, als wenn der Benutzer diesen Baum im Gedächtnis aufbaut oder auf Papier mitschreibt. Dennoch hat der Benutzer immer noch keine Hilfe, welcher der Knoten vielleicht die höchste Relvanz hätte. Die Benutzung einer Suchmaschine, d.h. eines Index, ist zwar möglich, steht und fällt aber mit der passenden Auswahl der zu suchenden Stichwörter.

Aufgabe der Erfindung ist es daher, eine Einrichtung anzugeben, die, ausgehend von einem bekannte Hypertext-Dokument, automatisch andere Dokumente anzeigt, ohne daß der Benutzer aus dem Inhalt des Ausgangsdokuments Suchworte extrahieren muß, um damit eine Index- oder Volltextsuche anzustoßen.

Darstellung der Erfindung

Die Erfindung verwendet die Erkenntnis, daß in vielen Fällen eine Seite ähnlichen Inhalts benötigt wird. Daher stellt die Erfindung eine Einrichtung bereit, mit der in einer symbolischen Darstellung eines Ausgangsdokuments und der damit verknüpften Dokumente zugleich mit dem Symbol der Grad der Ähnlichkeit zu dem ausgewählten Ausgangsdokument angezeigt wird.

Weitere Merkmale und Vorteile der Erfindung ergeben sich aus der folgenden Beschreibung, welche in Verbindung mit den beigefügten Zeichnungen die Erfindung an Hand eines Ausführungsbeispiels erläutert.

Kurzbeschreibung der Zeichnungen

Es zeigen

Fig. 1 ein von der Einrichtung angezeigtes Bild.

Beschreibung einer Ausführungsform der Erfindung

In der bevorzugten Ausführungsform besteht die Einrichtung aus einem Computer mit einer graphischen Anzeige und den bekannten Eingabeeinheiten wie Maus und Tastatur. Die graphische Anzeige wird bevorzugt mit den Programmpaketen X/Windows, JAVA, einem auf -ix endenden Betriebssystem usw. betrieben. Die Verwendung der häufig verkürzt als 'Windows' bezeichneten Programme der Firma Microsoft ist gleichfalls möglich.

Auf dieser Anzeige werde ein Dokument angezeigt, welches ein Hypertext-Dokument ist, das bevorzugt im HTML-Format gespeichert ist. Zu der Anzeige wird ein als Browser bezeichnetes Programm verwendet, welches die Formatanweisungen von HTML zur Darstellung auswertet. Für die vorliegende Erfindung sind dabei die Hypertext-Verweise von besonderer Bedeutung, im folgenden kurz als Verweise oder 'links' bezeichnet.

Durch beispielsweise eine JAVA-Anwendung wird dem Benutzer zusätzlich zu den ohnehin vom Browser bereitgestellten Funktionen eine Zusatzfunktion bereitgestellt, die im folgenden genauer beschrieben wird. Es ist aber auch ohne weiteres möglich, hierzu eine eigenes Programm in JAVA oder einer anderen geeigneten Programmiersprache zu verwenden, dem die als URL bezeichnete Adresse des Ausgangsdokuments als Parameter mitgegeben wird.

Dabei wird dieses Programm zunächst die in dem Ausgangsdokument enthaltenen Verweise benutzen, um zu den damit bezeichnenten Dokumeten zu gelangen, bei denen wiederum das Vorgehen rekursiv wiederholt wird. Da die Verweisstrukturen von Hypertext-Dokumenten nicht unbedingt einen Baum darstellen, ist eine Beschränkung der Suchtiefe notwendig. Diese erfolgt entweder durch die Angabe der Rekursionstiefe, z.B. vier, oder die Anzahl der besuchten Dokumente, oder die verbrauchte Zeit, oder eine Kombination hiervon. Auch kann festgelegt werden, daß nur Adressen einer bestimmten Domäne verfolgt werden.

In Fig. 1 ist eine im wesentliche baumartige Darstellung gezeigt, wie sie als Ergebnis eines rekursiven Abstiegs in vier Ebenen, das Ausgangsdokument eingeschlossen, nach dem Stand der Technik entstanden sein könnte. Die Dokumente sind als Kreise dargestellt und die Verweise als Pfeile. Die Schraffur einiger Kreise ist im Stand der Technik noch nicht vorhanden, sondern Teil der Erfindung. Offensichtlich enthält Dokument A1 zwei Verweise auf die Dokumente B1 und B2; B1 Verweise auf C1, C2, C3 und D4; B2 Verweise auf C3 und C4; C1 Verweise auf D1, D2 und D3; C2 Verweise auf D3, D4 und D5; C3 Verweise auf D5 und D6; C4 auf D7 und D8.

Da die Dokumente für die Bestimmung der in ihnen enthaltenen Verweise ohnehin in die Einrichtung geladen werden müssen, erfolgt für jedes neu geladene Dokument eine Bearbeitung. Hierbei werden die Wörter des Dokuments extrahiert und in ihrer Häufigkeit bewertet. Daß dabei nicht signifikate Wörter wie Artikel, Konjuktionen usw., sogenannte Stopwörter, ignoriert werden, versteht sich von selbst. Stark flektierende Sprachen sollten gegebenenfalls ein Wörterbuch o.ä. für die Bestimmung der Grundformen benutzen und dann nur die Grundformen verwenden.

Die Bewertung der Häufigkeit ist in erster Näherung einfach die Anzahl des Vorkommens in dem Dokument. In einer verbesssterten Variante wird berücksichtigt, wie der Ort des Vorkommmens markiert ist. Beispielsweise könnten Wörter im Titel bzw. der Stichwortliste mit höherem Gewicht bewertet werden, so daß die Häufigkeit als Bruch erscheint. Ferner ist eine Normierung auf die Gesamtzahl möglich. Damit ergibt sich ein mit der Anzahl der untersuchten Dokumente wachsende Matrix, in deren Zeilen die Dokumente und in deren Spalten die Wörter indiziert sind.

Mittels dieser Matrix kann durch Multiplikation zweier Zeilenvektoren und Summierung der Produkte ein Abstand zweier Dokumente bestimmt werden. Dieses Abstandsmaß ist umso größer, je ähnlicher sich die beiden Dokumente sind, weil die Zahl besonders groß ist, wenn die Dokumente gemeinsame Wörter haben, die zudem noch in beiden Dokumenten gleich häufig vorkommen. Die ersten Vorschläge in dieser Richtung wurden von H. Luhn in dem Artikel "The automatic creation of literature abstracts", IBM Journal of Research and Development 2, 158-165, 1958, vorgeschlagen. Andere Funktionen, die die Matrix verwenden oder aus der Matrix eine quadratische symmetrische Matrix des Abstands der Dokumente untereinander extrahieren, indem paarweise Abstände bestimmt und damit die Wörter eliminiert werden, sind gleichfalls möglich. Die Auswahl kann nach pragmatischen Gesichtspunkten erfolgen und ist ohne wesentliche Auswirkung auf die Grundfunktionalität der Erfindung, wenn auch davon ein praktischer Erfolg, bezogen auf ein Fachgebeit, wesentlich abhängen kann. Im übrigen entspricht das Abstandsmaß nicht den Kriterien eines topolgischen Abstands, da die Dreicksungleichung nicht erfüllt sein muß und der Abstand zu sich selbst einen maximalen Wert anstelle von Null liefert.

Die Verwendung von Worthäufigkeitsvektoren ist insofern vorteilhaft, als die Matrix der gewichteten Worthäufigkeiten dynamisch während des rekursiven Durchsuchens erfolgen kann und jedes Dokument nur einmal übertragen und analysiert werden muß. Dies schließt jedoch nicht aus, daß die Einrichtung auch derart betrieben wird, daß ein Abstandsmaß jedesmal neu bestimmt wird, indem die betroffenen Dokumente aktuell geladen und ausgewertet werden. Auch ist eine Kombination möglich, bei der die Bestimmung über Worthäufigkeiten eine Vorauswahl von Dokumenten bestimmt, für die dann paarweise das Abstandsmaß nach anderen Verfahren, die den Dokumententext selbst benötigen, genau bestimmt wird. Wie oben angedeutet, könnte dies für stark flektierende o.ä. Sprachen gelten, bei denen der Vorgang der Reduktion auf Wortstämme einer aufwendigen Syntax- und Semantikanalyste bedarf.

Bevorzugt wird, nachdem der Suchvorgang abgeschlossen und die Matrix erstellt ist, die Verweisstruktur angezeigt. Hierfür sind eine Vielzahl von Formen bekannt; beginnend bei einer Auflistung mit Einrückungen, einer baumähnlichen graphischen Darstellung oder aufwendigen 3D/2D Darstellungen. In allen üblichen Darstellungsformen steht eine Baumstruktur im Vordergrund, wie sie bei rekursiven Abstieg kanonisch entsteht. Die nicht der Baumstruktur entsprechenden Verweise werden dann entweder nicht gezeigt oder als zusätziche Linien, ggf. in schwacher Form, dargestellt. Als 3D/2D Darstellung sind verschiedene Formate bekannt, bei denen die Struktur zunächst als Graphik in einem dreidimensionalen Raum aufgebaut und dann auf eine zweidimensionale Fläche projeziert wird, wie sie z.B. als "Fisheye-View" bekannt ist.

In Fig. 1 ist eine stark vereinfache solche Darstellung zu sehen, bei der Farbe durch Schraffur dargestellt wird. Dokument A1 ist das Ausgangsdokument und besonders gekennzeichnet, hier durch eine doppelte Umrandung. Da es ferner der Ausgangspunkt der Ähnlichkeiten ist, hat es dieselbe Schraffur wie die beiden dazu ähnlichsten Dokumente D3 und C3. Die beiden nächst ähnlichen Dokumente B1 und D2 sind gepunktet dargestellt.

Unabhängig von der Darstellung besteht die Erfindung darin, daß der über die Matrix oder sonstwie bestimmte Abstand zu dem Ausgangsdokument durch die Symbole in der Strukturdarstellung angezeigt wird. Bevorzugt wird dabei Farbe verwendet, weil diese in den bekannten Darstellungen keine wesentliche Rolle spielt. Beispielsweise könnte Rot für die größte Ähnlichkeit, Grün für die nächstnähreren, über Gelb und Blau bis Schwarz für relative Unähnlichkeit verwendet werden. Graustufen stellen eine andere Art der Färbung da, wobei hier bei einer Anzeige mit hellem Hintergrund Weiß als wenig signifikant und schwarz als höchst ähnlich bevorzugt verwendet werden.

Ein Größe der Symbole ist gleichfalls einer Farbe äquivalent; daher steht in den Ansprüchen "Farbe" sowohl auch für Graustufen als auch für andere, skalierbare Darstellungen wie dem Durchmesser einer Kreisfläche. Lediglich bei den 3D/2D Darstellungen, bei dennen durch die Projektion eine perspektivische Verkleinerung gewünscht ist, um die Raumlage zu visualisieren, ist die Größe nicht als "Farbe" anwendbar. Die Verwendung der Form ist zwar auch möglich, weil ein Dreieck eine wesentlich siginfikantere Darstellung und deutlich von einem Quadarat unterscheidbar ist, wohingegen der Unterschied zwischen einem Sechs- und einem Siebeneck kaum sichtbar ist. Dennoch stellt in diesem Beispiel

die Eckenzahl auch eine "Farbe" dar. Für Benutzer mit reduzierter Sehfähigkeit bei Buntfarben, die meist durch bessere Unterscheidbarkeit von z.B. Formen kompensiert wird, ist diese Möglichkeit wichtig und kann mit der Buntfarbdarstellung kombiniert werden.

Sind die Abstände und Farben der Symbole bestimmt, dann ist noch ein besondere Hervorhebung der dem Ausgangsdokument am nächsten liegenden sinnvoll, beispielsweise durch ein ganz oder teilweise blinkendes Symbol, beispielsweise durch einen blinkenden gelben Umring, wenn die Symbole Kreisflächen sind und eine dunkle Farbe größere Ähnlichkeit signalisiert als eine helle.

Da die Symbole alle auf der Oberfläche erscheinen, kann auch ein Symbol, welches bislang nicht das Ausgangsdokument ist, durch ein Eingabegerät (Maus) zum neuen Ausgangsdokument gemacht werden. In der bevorzugten Ausführungsform mit bereits in Matrizen akkumulierten Daten kann dann schnell die neue Einfärbung des Darstellung bestimmt und angezeigt werden. Bevorzugt wird hierbei kein neuer Abstieg von der neuen Position aus durchgeführt, sondern es werden die bereits akkumulierten Daten verwendet. Bei entsprechenden Betriebsmitteln ist jedoch ein Hinzufügen der noch fehlenden, durch den neuen Bezugspunkt in Reichweite gerückten, Dokumente sinnvoll; gegebenenfalls als Hintergrundprozeß, der die Anzeige dann auf Anforderung auf den möglicherweise veränderten Stand bringt.

Da in der oben beschrieben Ausführung mit einer Matrixdarstellung von Wörtern und Dokumenten die Wörter noch als Liste verfügbar sind, können diese dem Benutzer als weiteres Auswahlmittel bereitgestellt werden. Hierzu ist eine alphabetische oder eine Sortierung nach Häufigkeit möglich. Wählt der Benutzer eines odere mehrere Wörter, so wird das-

jenige Dokument zum Ausgangsdokument, daß hierzu am besten paßt. Formal ist dann ein virtuelles Dokument Ausgangsdokument, das die maximalen Worthäufigkeiten der ausgewählten Wörter umfassen würde.

Im Unterschied zu einer Suche über einen Index wird dabei nicht eine geänderte graphische Anordnung der angezeigten Struktur bewirkt, sondern lediglich deren Einfärbung geändert.

Eine andere Ausführungsform verwendet, bevorzugt zusätzlich zur Farbe, noch den Abstand der Symbole im 3D-Raum zueinander als "Farbmerkmal". Gerade 3D-Darstellungen lassen noch erheblichen Spielraum im relativen Abstand der Symbole zueinander. Da jedoch die verwendeten Maße, wie oben angegeben, keine Metrik darstellen, ist ein solches Bild nicht eindeutig bestimmt. Durch iteratives Vorgehen kann jedoch eine Verformung bewirkt werden, die deutlich die relative Nähe verschiedener Dokumente anzeigt. Dabei kann in Kauf genommen werden, daß die Anzeige nicht stillsteht, sondern wegen der gegensätzlichen Einwirkungen die Anzeige sich ständig leicht verändert. Vielmehr ist dieses "Atmen" geeignet, die relative Unsicherheit der Einordnung besser anzuzeigen als ein "eingefrorenes" Bild, das eine finale Anordnung vortäuscht, die gar nicht stabil ist.

Patentansprüche

1. Einrichtung zum Suchen von bzw. Navigieren in durch Verweise verketteten Dokumenten, die auf einer Ausgabeeinheit symbolisch dargestellt sind,

   d a d u r c h g e k e n n z e i c h n e t ,

   daß, ausgehend von einem Ausgangsdokument, die Symbole der weiteren Dokumente mit einer Markierung versehen sind, die den Grad der Ähnlichkeit zu dem Ausgangsdokument gemäß einem Ähnlichkeitsmaß anzeigt.

2. Einrichtung nach Anspruch 1, wobei ein Ähnlichkeitsmaß verwendet wird, für das zunächst aus den Dokumenten ein Kennzahlvektor extrahiert wird, dieser in der Einrichtung abgelegt wird und sodann das Maß für die Änlichkeit zweier Dokumente ohne Rückgriff auf das jeweilige Dokument durch Operationen auf den Kennzahlvektoren bestimmt wird.

3. Einrichtung nach Anspruch 2, wobei in einem Vorbereitungsschritt, ausgehend von dem ersten Ausgangsdokument, den Verweisen bis zu einer vorgegeben Tiefe nachgegangen wird und von den besuchten Dokumenten der Kennzahlvektor extrahiert und abgelegt wird.

4. Einrichtung nach Anspruch 2 oder 3, wobei das Ähnlichkeitsmaß durch eine gewichtete Funktion über den Häufigkeiten gemeinsamer Wörter der zu vergleichenden Dokumente bestimmt wird.

5. Einrichtung nach einem der Ansprüche 1 bis 4, wobei der Grad der Änlichkeit durch die Gestaltung der Symbole angezeigt wird.

6. Einrichtung nach Anspurch 5, wobei als Mittel zur Gestaltung der Symbole Farbe verwendet wird.

7. Einrichtung nach Anspruch 5 oder 6, wobei alternativ oder zusäztlich zu der Gestaltung der Symbole selbst deren Anordnung zueinder so modifiziert wird, daß der Abstand zum Symbol des Ausgangsdokuments entsprechend der Abstandsmetrik bei großer Ähnlichkeit relativ gering wird.

8. Einrichtung nach einem der vorhergehenden Ansprüche, wobei ein Eingabegerät vorgesehen ist, mit dem eine die Einrichtung benutzende Person ein Symbol auswählen kann, welches dadurch das Ausgangsdokument bestimmt.

9. Einrichtung nach Anspruch 8 in Kombination mit einem der Ansprüche 4 bis 7, wobei eine Eingabevorrichtung vorgesehen ist, mittels derer eine die Einrichtung benutzende Person ein oder mehrere Wörter auswählen kann, womit das Dokument mit der größten gewichteteten Worthäufigkeit dieser Wörter zum Ausgangsdokument wird.

Zusammenfassung

Einrichtung zum Suchen von bzw. Navigieren in durch Verweise verketteten Dokumenten, die auf einer Ausgabeeinheit symbolisch dargestellt sind, wobei, ausgehend von einem Ausgangsdokument, die Symbole der weiteren Dokumente mit einer Markierung versehen sind, die den Grad der Ähnlichkeit zu dem Ausgangsdokument gemäß einem Ähnlichkeitsmaß anzeigt.

Fig. 1

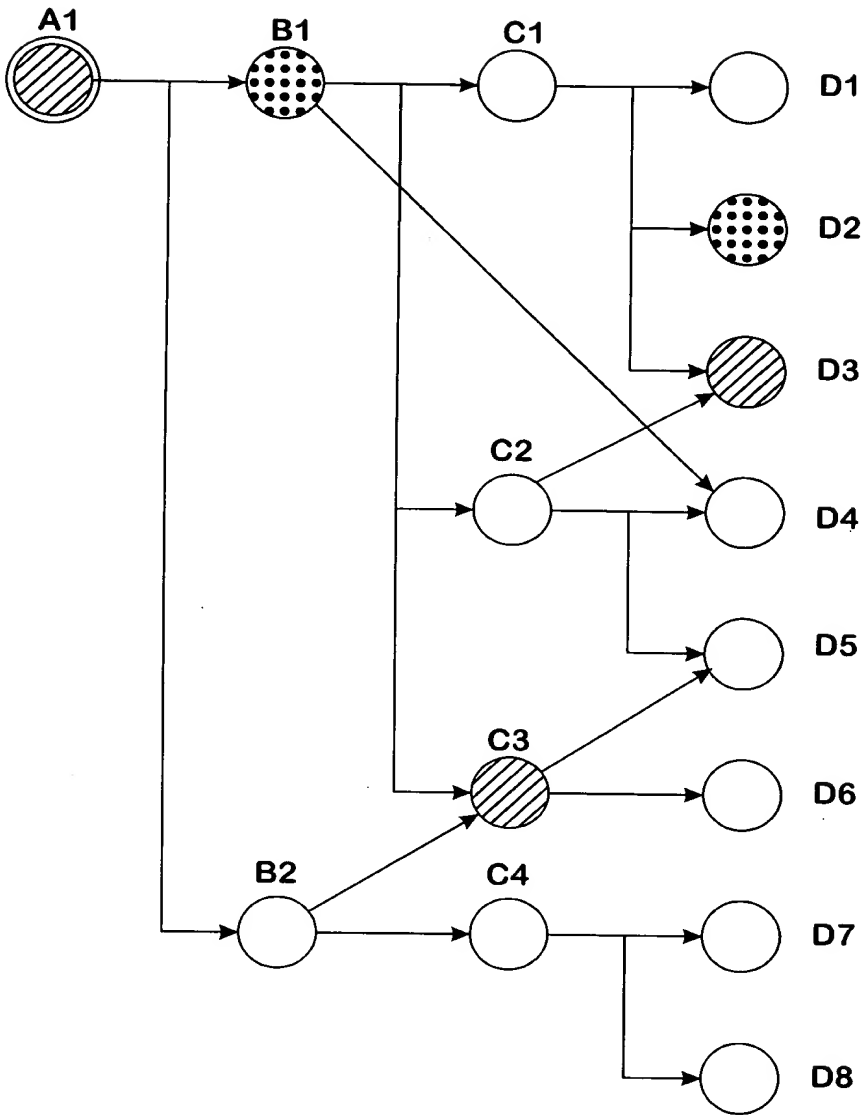

Fig. 1

Zusammenfassung